NATIONAL BUREAU OF STANDARDS REPORT

NBS PROJECT
1103-40-5150

NBS REPORT
4894

STATISTICAL ANALYSIS OF DATA ON CLASSIFICATION OF RADIOGRAPHS OF ALUMINUM ALLOY SPECIMENS

By

Joan R. Rosenblatt

Statistical Engineering Laboratory

Report to
Naval Ordnance Laboratory
Department of the Navy
Order No. 60921/5243/56

October 26, 1956

NBS

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

Report No.

Naval Ordnance Laboratory
Department of the Navy
Order No. 6092R/52Rj/56

STATISTICAL ANALYSIS OF DATA ON CLASSIFICATION OF RADIOGRAPHS OF ALUMINUM ALLOY SPECIMENS

For each of four aluminum alloys, specimens were classified by three readers according to the amount of a dispersed defect. Classification was made by comparison of radiographs with standard radiographs representing 1, 2, 3, etc., "degrees" of the defect.

In this report, the data on these classifications are analyzed to determine certain properties of the method of classification. The conclusions and summary tables are presented in six sections:

1) Definition of "Consensus"
2) Proportion of Scores Within $\pm$ 1 Degree of Consensus.
3) Differences Among Readers.
4) Variability of Interpretations.
5) Analysis of Repeated Readings.
6) Suggestions for Further Experiments.

1. Definition of "Consensus".

For the occasional cases where interpretation differences exceeded two degrees, the rule for consensus assignment followed in your worksheets did not seem to be well-defined.

The rule adopted by us, was to let the consensus be defined as the median of the three scores. Thus, if two or more observers agree the degree assignment given by them is the median or consensus. If each judge assigns a different

degree score to a radiograph, the middle number is the median. In the few cases where use of the median led to changing the value given for the consensus, the revised figure is noted (in purple) on the original data sheets.

2. Proportion of Scores Within ± 1 Degree of Consensus.

From the data which were obtained, one may calculate the proportion of cases in which a reader assigned a score within ± 1 degree of the consensus score of three readers. This may be interpreted as an estimate of the average proportion of assignments within ± 1 of consensus, which would be obtained in hypothetical repetitions of the same experiment (some readers, same set of radiographs). The extent to which this estimate can be used in relation to possible different sets of radiographs and different groups of readers depends on one's degree of belief that the collections of radiographs and the groups of readers involved in the present experiment are representative of the radiographs and readers about whom it is desired to make an inference or prediction.

From the estimated proportion and the number of readings made, one may calculate a lower confidence limit for the average proportion of readings within ± 1 of consensus. Thus, one may assert that (at the 0.975 probability level, say) the expected proportion of readings by Reader A within ± 1 of consensus is not less than a specified number. The confidence limits given in Table I are approximations based on the simplifying assumption that the probability of interpreting

within ± 1 of consensus is the same for each specimen of a given alloy. Table Ia gives the proportion of readings within ± 1 of consensus at each consensus value.

For reasons discussed below (Section 5), the first readings were used for this analysis. Revised scores were ignored.

TABLE I

Proportion of Scores Within ± 1 Degree of Consensus

Reader	Number of Readings	Proportion of Scores within ± 1 Degree of Consensus	0.975 Confidence Limit for Expected Proportion *
		195 Alloy	
A	422	1.000	.99
B	422	.967	.94
C	422	.922	.89
		355 Alloy	
A	447	.991	.98
B	449	1.000	.99
C	449	.998	.99
		Dow H (Transverse)	
A	371	.965	.94
B	371	.957	.93
C	371	.960	.93
		Dow H.(Longitudinal)	
A	327	.979	.96
B	327	.887	.85
C	327	.985	.96

*) Obtained from Table 41 in E. S. Pearson and H. O. Hartley (ed.), Biometrika Tables for Statisticians, Volume I, Cambridge University Press (1954).

TABLE Ia

Proportion of Scores Within ± 1 Degree of Consensus

Consensus	Number of Readings	Reader A	Reader B	Reader C
		195 Alloy		
1	66	1.00	1.00	.64
2	66	1.00	1.00	1.00
3	54	1.00	1.00	1.00
4	42	1.00	.86	1.00
5	38	1.00	.87	.95
6	47	1.00	1.00	1.00
7	46	1.00	.93	.89
8	63	1.00	1.00	1.00
		355 Alloy		
1	85	1.00	1.00	1:00
2	112*	.96	1.00	1.00
3	123	1.00	1.00	1.00
4	76	1.00	1.00	.99
5	49	1.00	1.00	1.00
6	4	1.00	1.00	1.00
		Dow H (Transverse)		
1	48	1.00	1:00	1.00
2	62	.97	1.00	.98
3	69	.93	.96	.94
4	82	.96	.94	.91
5	66	.98	.91	.97
6	44	1.00	.95	.98
		Dow H (Longitudinal)		
1	56	1.00	1.00	1.00
2	55	1.00	.98	1.00
3	54	.96	1.00	.98
4	56	.93	.91	.96
5	49	1.00	.82	.98
6	57	1.00	.61	.98

*) For Reader A there are only 110 readings.

TABLE Ia

Proportion of Scores Within ± 1 Degree of Consensus

Consensus	Number of Readings A	Reader B	C
	10% Alloy		
1	66	1.00	1.00
2	66	1.00	1.00
3	64	1.00	1.00
4	62	1.00	.86
5	50	1.00	.87
6	47	1.00	1.00
7	46	1.00	.93
8	63	1.00	1.00
	35% Alloy		
1	85	1.00	1.00
2	113*	.96	1.00
3	123	1.00	1.00
4	76	1.00	.93
5	49	1.00	1.00
6	44	1.00	1.00
	Dow H (Transverse)		
1	48	1.00	1.00
2	62	.97	.98
3	69	.93	.96
4	82	.96	.94
5	66	.98	.97
6	44	1.00	.98
	Dow H (Longitudinal)		
1	56	1.00	1.00
2	52	1.00	.98
3	54	.96	1.00
4	50	.92	.96
5	49	1.00	.82
6	37	1.00	.61

*For Reader A there are only 110 readings.

3. Differences Among Readers.

It must be said first that it is not possible to make an extensive analysis of observed differences among readers, since the experiment did not provide data permitting comparison of differences among readers with differences among repeated readings of the same radiograph by each reader.

One type of over-all picture of the differences among readers is given by Figures 1 through 4, which are based on Table II. Figure 1, for example, shows for each reader the distribution among the eight degrees of scores assigned to the 422 specimens of 195 Alloy; apparently Reader C may have a tendency to avoid assigning degrees 1 and 2 to radiographs of 195 Alloy specimens. Table II and Figures 1 through 4 are based on first readings.

In examining Figures 1 through 4, it must be remembered that the identity of a reader is not the same in all figures. the identities recorded on the worksheets are:

	Reader A	Reader B	Reader C
195 Alloy	Pierce	IJF	Criscuolo
355 Alloy	Polansky	IJF	Cris. or J D G
Dow H (Transverse)	Polansky	IJF	J D G
Dow H (Long.)	Polansky	IJF	J D G

These figures suggest that for some purposes the differences among interpretations of radiographs by different readers may be substantial. For example, if material is to be accepted or rejected according as the degree of defect (porosity or microshrinkage) is greater or less than a specified amount, it

could make quite a lot of difference if a reader had a marked tendency to avoid assigning scores at the endpoints of the scale.

It may be recorded that for every alloy the over-all distributions for the three readers are (statistically) significantly different. The same sort of conclusion is reached in other ways, e.g., for Dow H alloy Reader B assigns scores lower than Reader A for a significantly larger proportion of the radiographs on which they disagree. Whether these differences are practically meaningful or not must be determined in the context of the use which is to be made of the degree scores.

TABLE II

Distribution of Alloy Specimens According to Degree Scores Assigned by Individual Readers and by Consensus of Three Readers

Reader	Degree 1	2	3	4	5	6	7*	8*
				195 Alloy				
A	72	81	39	37	37	51	42	63
B	65	63	51	33	42	40	62	66
C	26	50	105	59	47	46	29	60
Consensus	66	66	54	42	38	47	46	63
				355 Alloy				
A	86	106	106	77	51	21	-	-
B**	85	99	141	75	47	0	-	-
C**	85	120	114	75	46	7	-	-
Consensus	85	110	123	76	49	4	-	-
				Dow H (Transverse)				
A	43	54	60	74	69	71	-	-
B	53	73	87	72	51	35	-	-
C	56	61	53	84	59	58	-	-
Consensus	48	62	69	82	66	44	-	-
				Dow H (Longitudinal)				
A	50	49	59	45	46	78	-	-
B	54	72	72	79	29	21	-	-
C	58	53	46	61	48	61	-	-
Consensus	56	55	54	56	49	57	-	-

*) Degrees 7 and 8 are used only for 195 Alloy.

**) There are two lines in the worksheets where readings are missing for Reader A. These two cases are omitted in this tabulation, so that each row shows the distribution of 447 readings.

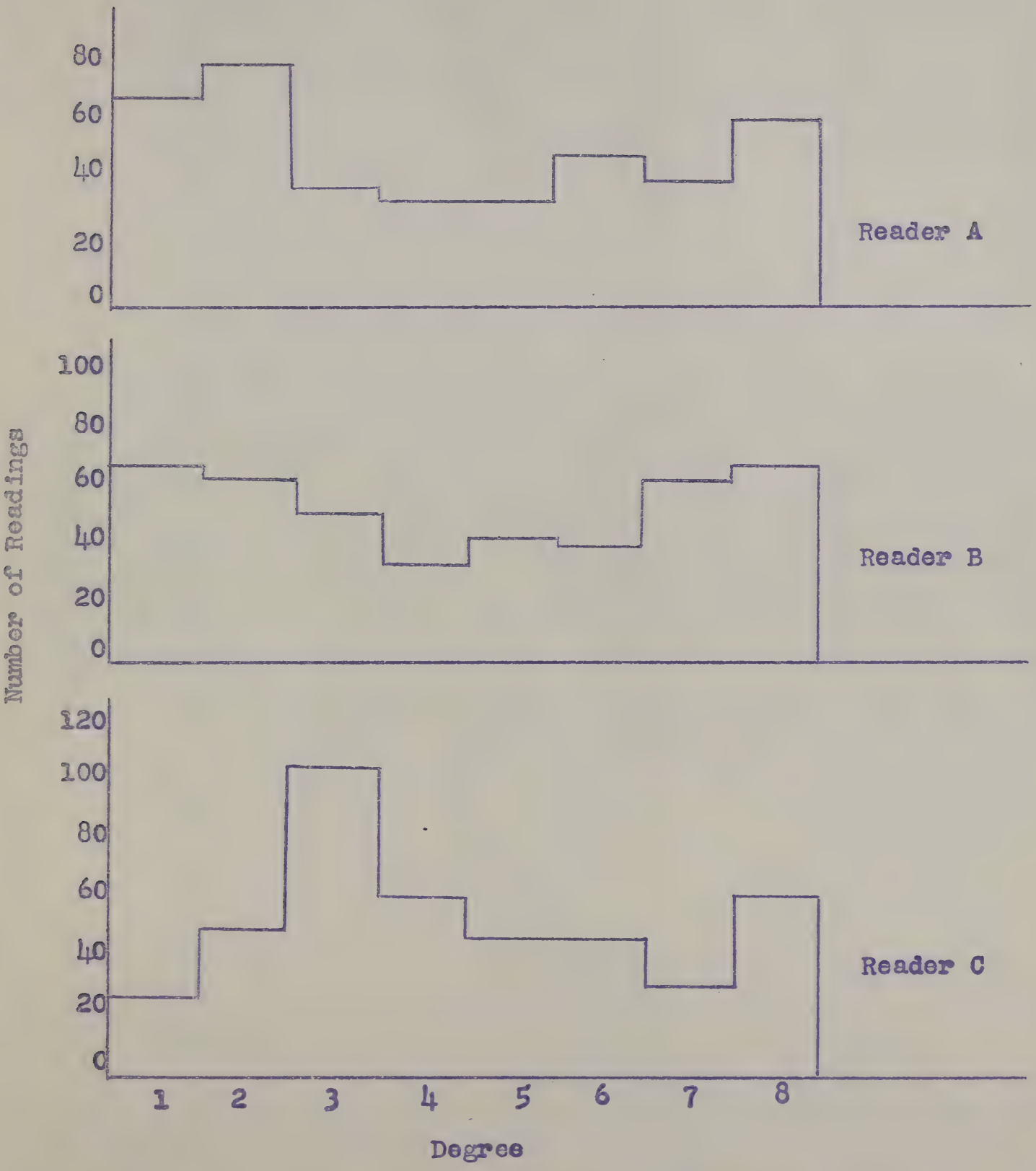


Figure 1. Distribution of Readings, 195 Alloy

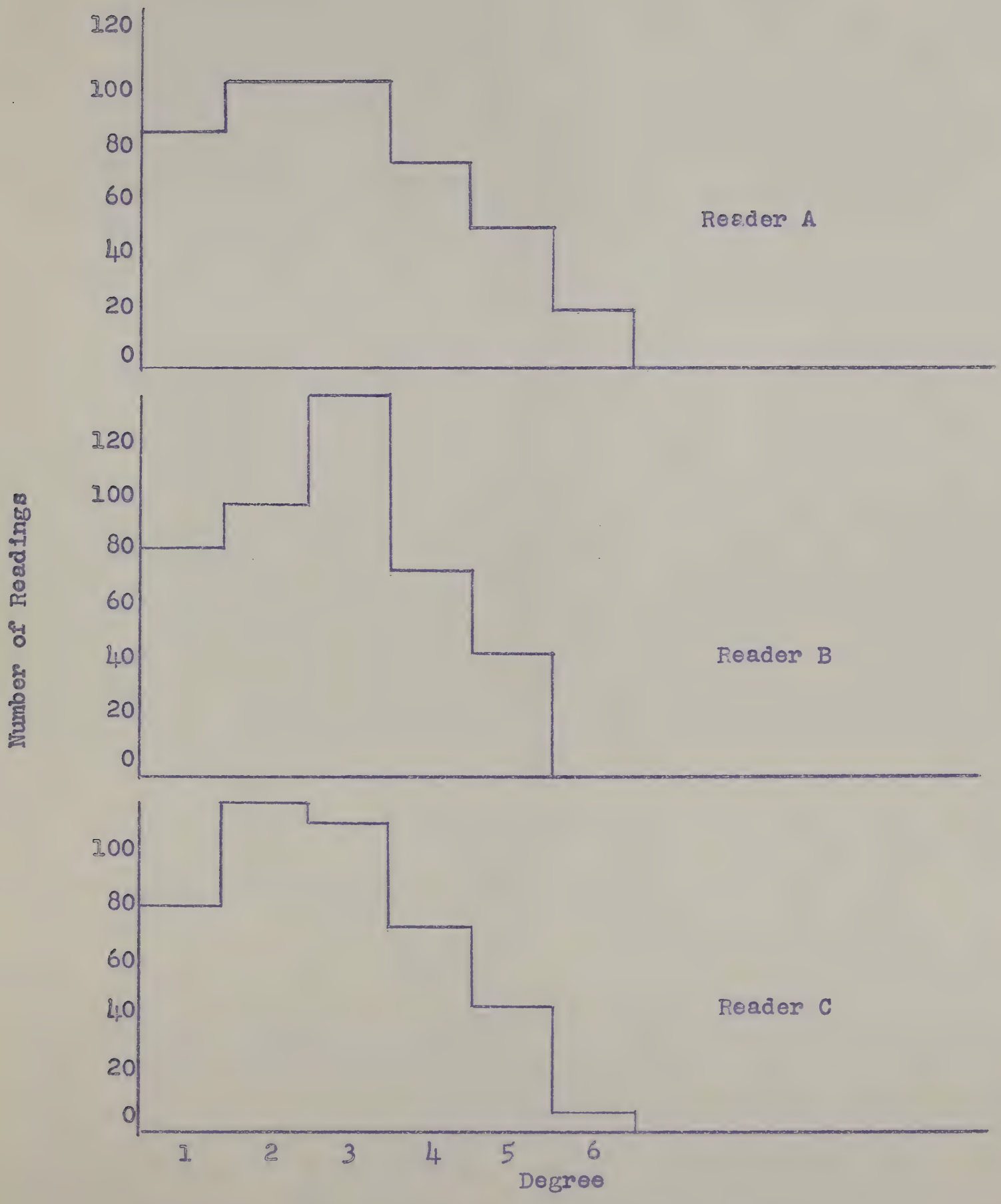


Figure 2. Distribution of Readings, 355 Alloy

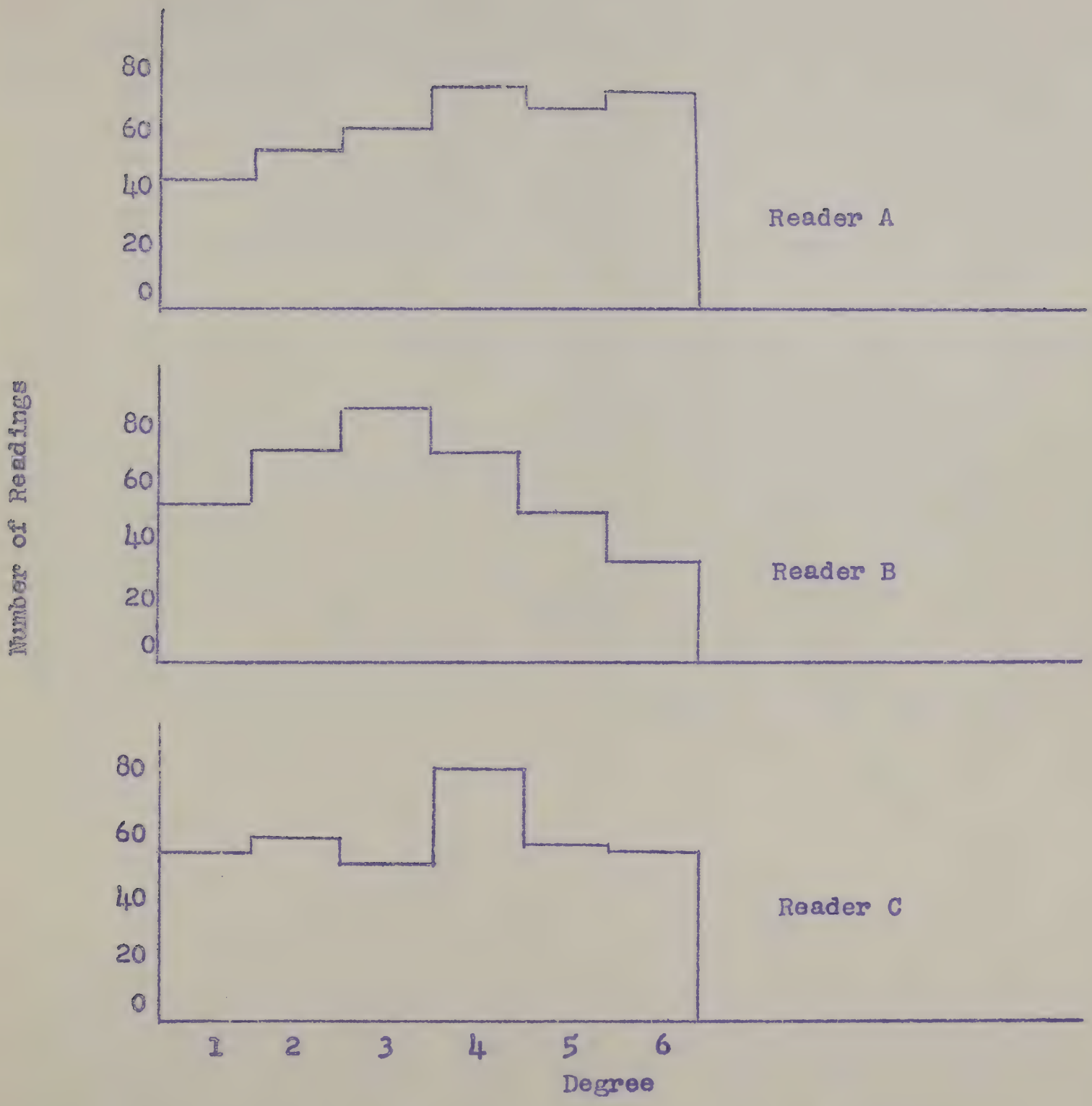
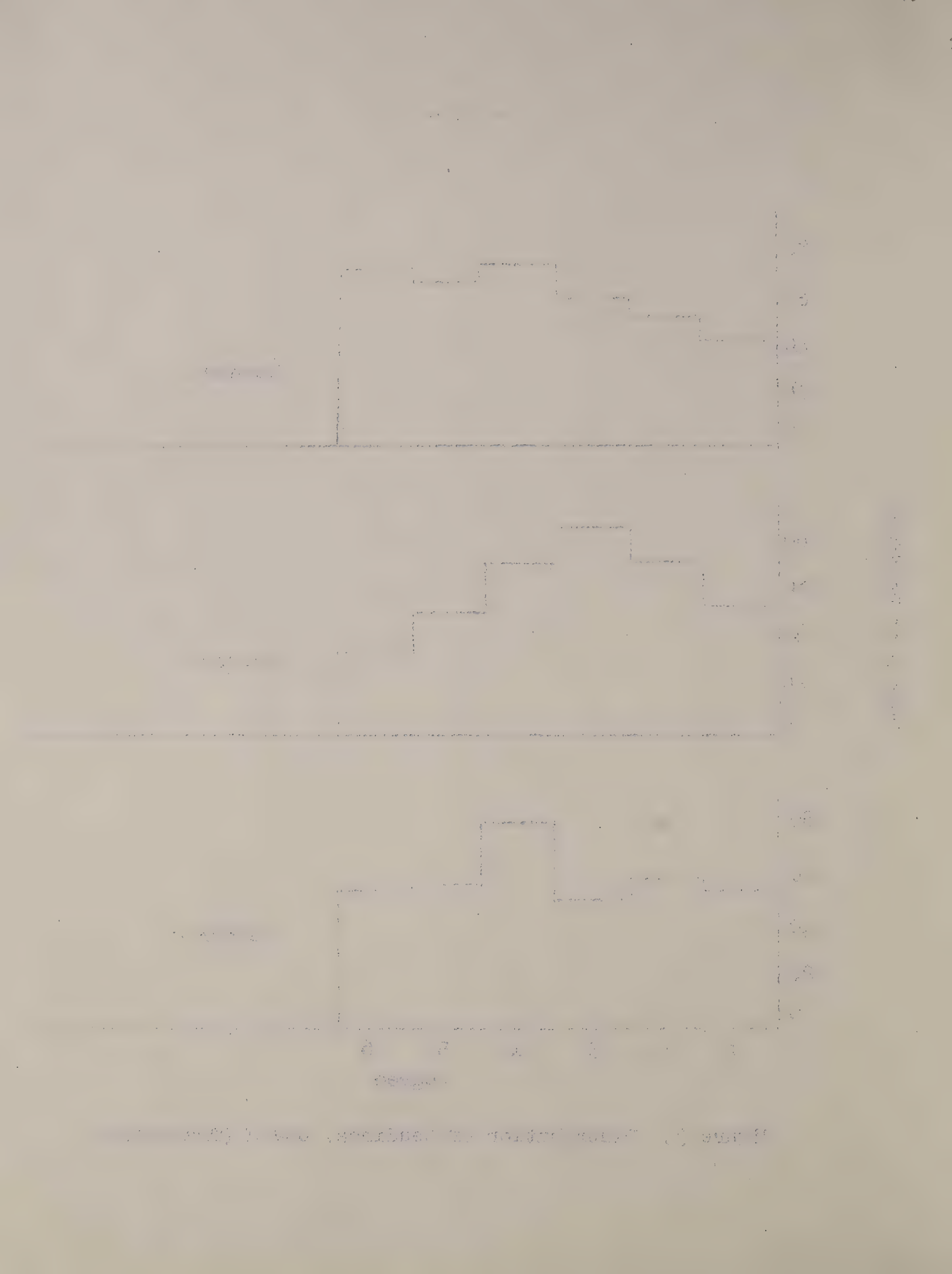


Figure 3. Distribution of Readings, Dow H (Transverse)



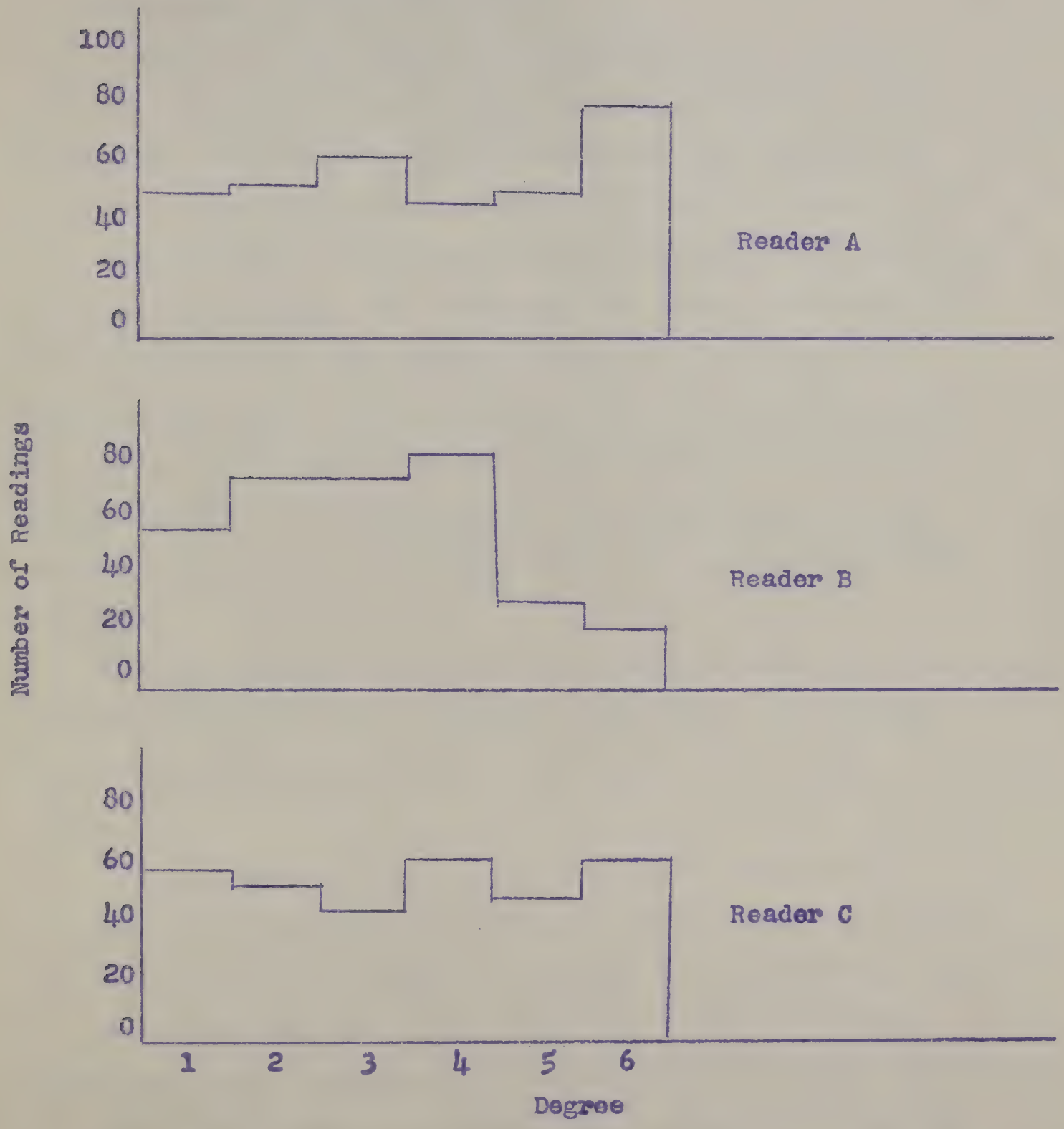


Figure 4. Distribution of Readings, Dow H (Longitudinal)

4. Variability of Interpretations.

A measure of the amount of variability among readers of the same radiograph is the range of scores assigned to that radiograph, that is, the difference between the largest and smallest of the three scores. Table III shows for each degree the distribution of these ranges for specimens having consensus-score at that degree. To summarize, the table also shows the proportion of cases for which the range was less than or equal to one, that is, the proportion of cases in which either all three readers agreed or two readers agreed and the third reader assigned an adjacent score. The distribution for all readers combined is also shown. Table III is based on first readings.

The data provide no way to distinguish between the two possible sources of this observed variability: differences among readers and variability of interpretations by one reader.

A third factor affecting variability of interpretation is that for radiographs exhibiting an extreme amount of defect (at either end of the scale) the readers will necessarily tend to give scores that are closer together, since the scoring method not permit assignment of scores beyond the range represented by comparison standards. The effect of this factor is seen in Table III, where it is evident that the wider ranges of scores tend to occur more frequently when the

consensus is nearer the middle of the scale.

Another measure of variability, which is more appropriate for comparing sources of variability, is the mean difference (see sections 5 and 6 below).

TABLE III

Distribution of Range of Scores Assigned by Three Observers

Consensus	Number of Readings	Range of Scores 0	1	2	3	4	Proportion of Readings with Range ≤ 1
			195 Alloy				
1	66	23	19	22	2	-	.64
2	66	16	46	4	-	-	.94
3	54	10	38	6	-	-	.89
4	42	15	19	2	6	-	.81
5	38	12	17	7	2	-	.76
6	47	9	32	6	-	-	.87
7	46	9	25	12	-	-	.74
8	63	54	9	-	-	-	1.00
Total	422	148	205	59	10	0	.84
			355 Alloy				
1	85	85	-	-	-	-	1.00
2	110*	69	37	4	-	-	.96
3	123	52	69	2	-	-	.98
4	76	24	47	5	-	-	.93
5	49	16	33	-	-	-	1.00
6	4	-	4	-	-	-	1.00
Total	447	246	190	11	0	0	.98
			Dow H (Transverse)				
1	48	37	11	-	-	-	1.00
2	62	21	34	6	1	-	.89
3	69	11	40	12	4	2	.74
4	82	21	32	21	6	2	.65
5	66	12	31	19	4	-	.65
6	44	26	15	3	-	-	.93
Total	371	128	163	61	15	4	.78

*) Two cases for which there were only two readings are omitted, since this is a tabulation of ranges of scores by three observers.

TABLE III (Continued)

Consensus	Number of Readings	Range of Scores 0	1	2	3	4	Proportion of Readings with Range ≤ 1
		Dow H (Longitudinal)					
1	56	36	20	-	-	-	1.00
2	55	16	38	1	-	-	.98
3	54	10	35	9	-	-	.83
4	56	12	27	12	4	1	.70
5	49	3	18	25	2	1	.43
6	57	14	20	20	3	-	.60
Total	327	91	158	67	9	2	.76

5. Analysis of Repeated Readings.

For the Dow H alloy, repeated readings were obtained in most cases where the range of the three first readings was two or more. The reasons for using the first readings in the foregoing analyses are the following.

First, revised readings were obtained only in cases where the range of first readings was large. Suppose, however, that the chance of a wide range among the three scores were about the same for every radiograph. Then one would expect that the ranges of the repeated readings would tend to be smaller than the ranges of first readings for the same radiographs if only the wide-range cases were repeated. For if the second readings were independent of the first, one would expect to find the distribution of ranges for second readings similar to the distribution of ranges among all first readings. Table IV indicates

that these expectations are roughly realized. Accordingly, use of the revised readings might lead to under-estimation of the amount of variability inherent in the process of interpreting radiographs, and to overestimation of the proportion of readings within ± 1 degree of consensus. The amount of difference that is involved may be illustrated by the following quantities calculated for Dow H (Transverse). In calculating the second column, first readings were replaced by second readings wherever available.

	First Readings	Revised Readings
Proportion Within ± 1 of Consensus		
A	.96	.99
B	.96	.98
C	.96	.98
Proportion of Ranges ≤ 1	.78	.88

The tendency for the ranges of second readings to be less than those of the first readings is seen even more sharply as follows. The difference between first and second ranges for each radiograph is positive (i.e., the second range is smaller) in an overwhelming number of cases.

	First Range Minus Second Range					
Dow H Alloy	-1	0	1	2	3	4
Transverse	-	10	25	16	3	1
Longitudinal	2	23	33	10	1	-

Second, the revised readings may not be strictly comparable to first readings. The readers knew when making

second readings that their first readings had had a wide range. They may in some cases have been able to recall the score assigned in the first reading, and the direction (below or above) in which this first score deviated from the median of first readings.

The repeated readings might have afforded data for comparison of variability among readings by one reader with differences among readers. The two considerations discussed above, however, make this impossible. Both have the effect of providing data on variability between readings by one reader under conditions which lead to greatest variability.

The mean (absolute) difference is appropriate in this context as a measure of variability. The mean difference among readers is calculated by summing the three absolute differences for each radiograph, and dividing the total by three times the number of radiographs. It may be recorded that the mean differences among readers for first readings were as follows.

Alloy	Mean Difference Among Readers
195	.57
355	.32
Dow H (trans.)	.62
Dow H (long.)	.67

The mean difference among readings is calculated by summing for each reader the absolute differences between his first and second readings and dividing by the number of cases for which there were two readings.

For Dow H Alloy, the mean differences between first and second readings of the same radiograph were as follows:

Reader	Mean Difference Between Readings A	B	C	Combined
Transverse	.84	1.25	.95	1.01
Longitudinal	.83	.61	.48	.64

The data on repeated readings might be used to provide information on the variability of the consensus, were it not again for the difficulties mentioned above. For example, we may compare the mean (absolute) difference between consensuses of first and second readings with the mean difference between individual readers' first and second readings.

Dow H Alloy	Mean Difference Between Consensuses	Between Readings
Transverse	.82	1.01
Longitudinal	.43	.64

As expected, the consensus is less variable than individual readings. For the reasons noted above, however, these are undoubtedly over-estimates of the variability inherent in the interpretation process.

The "reliability" of the consensus may be represented by the proportion of cases in which the consensus of first readings and the consensus of second readings differ by at most one degree. This is found to be 0.84 for Dow H (transverse) and 0.93 for Dow H (longitudinal).

In a few cases, a third reading was obtained. There are not enough such cases to permit satisfactory statistical analysis.

TABLE IV

Distribution of Ranges of Three Readings, for First and Second Readings on the Same Radiographs, Compared With Distribution of Ranges for All First Readings

Set of Readings	Number of Readings	Range 0	1	2	3	4
		Dow H (Transverse)				
First*	55**	-	-	37	14	4
Second	55	12	24	18	1	-
All First	371	128	163	61	15	4
		Dow H (Longitudinal)				
First*	69	-	-	58	9	2
Second	69	7	31	27	4	-
All First	327	91	158	67	9	2

*) First readings for radiographs for which there is also a second reading.

**) One case where only two readers made second readings is omitted.

6. Suggestions for Further Experiments.

In this section are noted, first, some possibilities for additional experiments which would contribute to more satisfactory analysis of the present experiment; and second,

further possibilities for experiments which might be useful if it were desired to answer certain additional questions about the process of radiograph interpretation.

It would be highly desirable to obtain independent second readings of at least some of the radiographs used in the present experiment, by the same readers. Such data are essential if it is desired to make a valid statistical statement about the variability of the consensus. Such data would also provide a basis for deciding whether the differences among readers are important relative to variability among successive readings by one reader. Independent second readings would be obtained by repeating the experimental process for the whole set of radiographs or for an appropriately selected random sample, so that obtaining a second reading does not depend on the outcome of the first readings. Furthermore, the readers should, while making the second reading, have no way of knowing what the outcome of the first reading was. In an experiment where first and second readings are made within a short time interval, precautions must be taken to meet this latter condition as nearly as possible.

A completed analysis of this experiment would then provide estimates of

(1) Expected proportion of readings within $\pm$ degree of consensus.

(2) Expected proportion of cases in which the consensus of two independent sets of readings differ by at most one degree.

(3) Mean difference between readers.

(4) Mean difference between two independent readings by one reader.

One would also obtain an idea of the relative magnitude of the two main sources of variability, reader differences and uncertainty of interpretation. If reader differences appear to be important, one could further determine what the biases of individual readers are (e.g., tendency to avoid ends of scale, tendency to assign lower scores).

Representativeness of radiographs.

A new experiment conducted in the same fashion as the present one (with independent repeated readings), but using some of the present radiographs and some new ones, would provide information as to the possibility of generalizing the application of the results of this experiment.

Representativeness of readers.

Similarly, the above experiment might be expanded by being duplicated with three new readers.

Effect of extreme points of scale.

An experiment might be conducted with different rules, namely, with the possibility of assigning an extreme specimen to a "degree" beyond the end of the scale.